A hybrid solution to improve iteration efficiency in the distributed learning

Junxiong Wang chuangzhetianxia@gmail.com
Harbin Institute of Technology, Harbin, China

Hongzhi Wang wangzh@hit.edu.cn
Harbin Institute of Technology, Harbin, China

Chenxu Zhao zhaochenxu001@gmail.com
Harbin Institute of Technology, Harbin, China

Abstract

Currently, many machine learning algorithms contain lots of iterations. When it comes to existing large-scale distributed systems, some slave nodes may break down or have lower efficiency. Therefore traditional machine learning algorithm may fail because of the instability of distributed system.We presents a hybrid approach which not only own a high fault-tolerant but also achieve a balance of performance and efficiency.For each iteration, the result of slow machines will be abandoned. First, we discuss the relationship between accuracy and abandon rate. Then we debate the convergence speed of this process. Finally, our experiments demonstrate our idea can dramatically reduce calculation time and be used in many platforms.

1. Introduction

With an explosion in the number of massive-scale data, the cost time of traditional solutions is unacceptable. As a result of it, the demand for scalable disturbed platforms and frameworks is rising in many areas including machine learning, data mining and pattern recognition.

Hadoop is widely used to speed up many machine learning algorithms applied in academy and industry. However, the serve problem about it is the lack of support for iterative programs. Thus lots of popular-used platforms and distributed frameworks appeared to optimize the iteration process for higher efficiency and less cost. A framework named Spark[1], widely used with a library for machine learning (MLlib), has improved the efficiency in many applications. Its modified solution, Resilient Distributed Datasets[2], speed up the data storage to an amazing step. In addition, Haloop[3] makes lots of improvement on iteration. It provides some rules to consult when developing the machine learning project based on map/reduce.

When it comes to applications, some slave nodes, taking up only a small percentage of all, always cost much more time than others in one iteration because of communication fault or their low efficiency. Traditional solutions cannot handle it as they have to calculate it again when failure occurs.

This paper presents a novel algorithm which is able to make a suitable balance in performance and efficiency. The master node doesn't have to wait all the slave nodes. After a certain percentage of slaves has finished calculated and communicated with the master node, the master one will start next iteration rather than waiting others. This process can not only improve the efficiency dramatically but also have a high fault tolerance because some nodes' fault do not have influence on this system. This algorithm is developed to decrease the total time, combining the synchronous and asynchronous process, resulting in reasonable efficiency and accuracy.

We discuss the relationship between the abandon rate and the accuracy with statistic. In next section, this algorithm is proved to converge and the speed of convergence is Q-linear convergence with the mathematics proof. It is such an excepted result that balance the efficiency and performance.

This idea can be applied to a list of algorithms including iterations such as Stochastic Gradient Descent, Conjugate Gradient Descent, L-BFGS and so on. Many existing platforms and framework can be improved with

this approach.

2. Algorithm

When our approach is applied in Gradient Descent, we give the following algorithm.

Suppose machine num is M and examples in each machine are ζ. For each iteration, the number of slaves that master should wait is γ. That is to say, only $\gamma\zeta$ examples can be calculated.

Estimating the least number of slave nodes which need to communicate with master node is the first step.

Algorithm 1 Calculate The Least Number of Slave Nodes

Input: $The\ capacity\ of\ data : N$
$Confidence\ coefficient : \alpha$
$Relative\ error\ \xi$
$Examples\ in\ each\ machine : \zeta$
Output: $Estimated\ Machine\ Number : \gamma$

$$\gamma = \frac{Nu_{\alpha/2}^2}{(\xi^2 N + u_{\alpha/2}^2) * \zeta};$$

Algorithm 2 Master's algorithm

1: **while** IsConvergence == false **do**
2: **if** received γ slave nodes **then**
3: $\theta_{i+1} = \theta_i - \frac{\eta_t}{\gamma} \sum_{j=1}^{\gamma} \theta_i^j$
4: **end if**
5: **end while**

Algorithm 3 Slaves' algorithm

1: Receive(θ^t)
2: $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \left\{ \frac{1}{\zeta} \sum_{i=1}^{\zeta} (\boldsymbol{\theta}^T \boldsymbol{K}[x_i] - y_i) \boldsymbol{K}[x_i] + \lambda \boldsymbol{\theta}^t \right\}$
3: Send(θ^{t+1})

3. The Algorithm's Convergence

In this section we will prove the convergence of this algorithm with an example of the quadratic programming. A definition of quadratic programming problem will be given first and then we will prove the convergence to explain the algorithm's correctness.At the end of this section, we will discuss the speed of convergence.

3.1. Brief Introduction

We donate $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{Y} \in \mathbb{R}^m$ as the example input and output space respectively. And set $\boldsymbol{\theta^*} \in \mathbb{R}^l$ as the optimal solution to a quadratic programming, then

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^l} \frac{1}{m} \sum_{i=1}^{m} (f_{\boldsymbol{\theta}}(x_i) - y_i)^2 + \lambda \left\| \boldsymbol{\theta} \right\|_{l^2}^2 \tag{1}$$

where $\lambda > 0$ is the regularization parameter.

Definition 3.1. *Denote function $\boldsymbol{K}$ is a kernel function where* $\boldsymbol{K}[x_i] = \begin{bmatrix} x_{i1}^2 \\ x_{i1}x_{i2} \\ \vdots \\ x_{i2}^2 \\ x_{i2}x_{i3} \\ \vdots \\ x_{in}^2 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \\ 1 \end{bmatrix} \in \mathbb{R}^l.$

Above all, $\boldsymbol{\theta^*}$ can be written as

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^l} \frac{1}{m} \sum_{i=1}^{m} (\boldsymbol{\theta}^T \boldsymbol{K}[x_i] - y_i)^2 + \lambda \left\| \boldsymbol{\theta} \right\|_{l^2}^2 \tag{2}$$

In the t th iteration,$\boldsymbol{\theta}^t$ can be updated as

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta_t \{ \frac{1}{\omega} \sum_{i=1}^{\omega} (\boldsymbol{\theta}^T \boldsymbol{K}[x_i] - y_i) \boldsymbol{K}[x_i] + \lambda \boldsymbol{\theta}^t \} \tag{3}$$

3.2. The Proof of Convergence

According to Taylor formula, (3) can turn into

$$\begin{aligned} f(\boldsymbol{\theta}^{t+1}) = f(\boldsymbol{\theta}^t) - \\ \eta_t \nabla f(\boldsymbol{\theta}^t)^T \{ \frac{1}{\omega} \sum_{i=1}^{\omega} (\boldsymbol{\theta}^T \boldsymbol{K}[x_i] - y_i) \boldsymbol{K}[x_i] + \lambda \boldsymbol{\theta}^t \} \\ + \mathrm{O}(0) \end{aligned} \tag{4}$$

If $f(\boldsymbol{\theta}^{t+1}) - f(\boldsymbol{\theta}^t) < 0$, then

$$\begin{aligned} \eta_t \{ \frac{1}{m} \sum_{i=1}^{m} (\boldsymbol{\theta}^T \boldsymbol{K}[x_i] - y_i) \boldsymbol{K}[x_i] + \lambda \boldsymbol{\theta}^t \}^T \\ * \{ \frac{1}{\omega} \sum_{i=1}^{\omega} (\boldsymbol{\theta}^T \boldsymbol{K}[x_i] - y_i) \boldsymbol{K}[x_i] + \lambda \boldsymbol{\theta}^t \} > 0 \end{aligned} \tag{5}$$

where the step size $\eta_t > 0$.
In order to prove the correctness of (5),we mention some useful lemma.

Lemma 3.1. *Donate the overall variance is σ^2,we choose a random sample of n (sampled without repeating) capacity from N elements. The variance of the sample mean is* $\sigma_n = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

Then, we proof this by statics theory. We define Z and z stand for the overall collection and sample collection respectively. So,

$$\sigma_n = \sqrt{\frac{1}{C_N^n}\sum_{i=1}^{c_N^n}(\overline{z_i} - \overline{Z})} \tag{6}$$

Then,

$$\begin{aligned}
\sigma_n^2 &= \frac{1}{C_N^n}\sum_{i=1}^{C_N^n}[\frac{(z_{i1}+z_{i2}+\cdots+z_{in})}{n} - \overline{Z}]^2 \\
&= \frac{1}{C_N^n n^2}\sum_{i=1}^{C_N^n}[(z_{i1}+z_{i2}+\cdots+z_{in}) - n\overline{Z}]^2 \\
&= \frac{1}{C_N^n n^2}[\sum_{i=1}^{C_N^n}(z_{i1}+z_{i2}+\cdots+z_{in})^2 \\
&- 2n\overline{Z}\sum_{i=1}^{C_N^n}(z_{i1}+z_{i2}+\cdots+z_{in}) + n^2 C_N^n \overline{Z}^2] \\
&= \frac{1}{C_N^n n^2}[\sum_{i=1}^{C_N^n}(z_{i1}+z_{i2}+\cdots+z_{in})^2 \\
&- 2n\overline{Z}C_{N-1}^{n-1}\sum_{S=1}^{N} Z_S + n^2 C_N^n \overline{Z}^2] \\
&= \frac{1}{C_N^n n^2}\sum_{i=1}^{C_N^n}(z_{i1}+z_{i2}+\cdots+z_{in})^2 - \overline{Z}^2
\end{aligned} \tag{7}$$

Every sample has a Z_S, then the expansion of the sample sum square has one Z_S^2 , so there are C_{N-1}^{n-1} Z_S^2 of $\sum_{i=1}^{M}(z_{i1}+z_{i2}+\cdots+z_{in})$, the total number of quadratic term is $C_{N-1}^{n-1}N = nC_N^n$. There are $(n^2-n)C_N^n$ $Z_iZ_j(i \neq j)$ and $\frac{N(N-1)}{2}$ kinds of Z_iZ_j. So every kind has $\frac{2n(n-1)C_N^n}{N(N-1)}$ terms.

$$\begin{aligned}
&\frac{1}{C_N^n n^2}\sum_{i=1}^{C_N^n}(z_{i1}+z_{i2}+\cdots+z_{in})^2 - \overline{Z}^2 \\
=&\frac{1}{C_N^n n^2}\{C_{N-1}^{n-1}\sum_{S=1}^{N} Z_S^2 + \frac{2n(n-1)C_N^n}{N(N-1)}\sum_{i<j} Z_iZ_j\} - \overline{Z}^2 \\
=&(\frac{1}{nN} - \frac{1}{N^2})\sum_{S=1}^{M} Z_S^2 + 2(\frac{n-1}{N(N-1)} - \frac{1}{N^2})\sum_{i<j} Z_iZ_j \\
=&(\frac{N-n}{n(N-1)})(\frac{N-1}{N^2}\sum_{S=1}^{N} Z_S^2 - \frac{2}{N^2}\sum_{i<j} Z_iZ_j) \\
=&(\frac{N-n}{n(N-1)})[\frac{1}{N}\sum_{S=1}^{N} Z_S^2 - \frac{1}{N^2}(\sum_{S=1}^{N} + 2\sum_{i<j} Z_iZ_j)] \\
=&\frac{N-n}{N-1}\frac{\sigma^2}{n}
\end{aligned} \tag{8}$$

Above all, Lemma 2.1 is proved to be right.

Lemma 3.2. *If the sample size* $n >= \frac{Nu_{\alpha/2}^2 s^2}{\Delta^2 N + u_{\alpha/2}^2 s^2}$, *the confidence coefficient of error under* $1-\Delta$ *is* α

Proof. If confidence coefficient is $1-\Delta$. Then,

$$P[\mid \overline{z} - \overline{Z} \mid < \Delta] = 1 - \alpha \tag{9}$$

When n is large,we can utilize normal approximation to make a conclusion that

$$P\{\mid \frac{\overline{z} - \overline{Z}}{\sigma_{\overline{z}}} \mid < \frac{\Delta_{\overline{z}}}{\sigma_{\overline{z}}}\} = \Phi(u_{a/2}) - \Phi(-u_{a/2}) \tag{10}$$

Then,

$$\Delta_{\overline{z}} = u_{a/2}\sigma_{\overline{z}} \tag{11}$$

Combining Lemma 2.1 ,we know

$$\Delta_{\overline{z}} = u_{a/2}\frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}} \tag{12}$$

$$n = \frac{Nu_{a/2}^2 s^2}{\Delta_{\overline{z}}^2 N + u_{a/2}^2 s^2} \tag{13}$$

Above all, the correctness of Lemma 2.2 can be proved. □

Using these Lemma, we can debate some questions about (5).Denote a set named Z

$$\begin{aligned}
Z = \{&(\boldsymbol{\theta}^T\boldsymbol{K}[x_1] - y_1)\boldsymbol{K}[x_1], \\
&(\boldsymbol{\theta}^T\boldsymbol{K}[x_2] - y_2)\boldsymbol{K}[x_2], \cdots, \\
&(\boldsymbol{\theta}^T\boldsymbol{K}[x_m] - y_m)\boldsymbol{K}[x_m]\}
\end{aligned} \tag{14}$$

Take ω elements from the collection Z, if we donate the average num of these elements is $\overline{\omega}$.
When $\Delta = \left|\xi\overline{Z}\right|$ is small, the correctness of (5) can be guaranteed. Combining (14), we can get $n = \dfrac{Nu_{a/2}^2 s^2}{(\xi\overline{Z})^2 N + u_{a/2}^2 s^2} \leq \dfrac{Nu_{a/2}^2 s^2}{\xi^2 s^2 N + u_{a/2}^2 s^2} = \dfrac{Nu_{a/2}^2}{\xi^2 N + u_{a/2}^2}$

3.3. Speed of Convergence

Definition 3.2. *Suppose that this process produce a sequence of iterations*$(\boldsymbol{\theta}^t)$ *converge to* $(\boldsymbol{\theta}^*)$*, if there exist a real number* β > *0 and a constant(q* > *0) which has no relationship with the number of iterations(t) let* $\lim_{t\to\infty} \dfrac{\left\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^*\right\|_{l^2}}{\left\|\boldsymbol{\theta}^k - \boldsymbol{\theta}^*\right\|_{l_2}^{\beta}} = q$*. Therefore, the sequence* $\boldsymbol{\theta}^t$ *has a convergence speed of Q-*β *-th.*

Our algorithm has a linear convergence of Q-th. That's to say β = 1 and q > 0.
We denote $B^t = \dfrac{1}{\omega}\sum_{i=1}^{\omega}(\boldsymbol{\theta}^T\boldsymbol{K}[x_i] - y_i)\boldsymbol{K}[x_i] + \lambda\boldsymbol{\theta}^t$. Then

$$\begin{aligned}
&\left\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^*\right\|_{l^2}^2 \\
&= \left\langle \boldsymbol{\theta}^t - \boldsymbol{\theta}^* - \eta_t B_t, \boldsymbol{\theta}^t - \boldsymbol{\theta}^* - \eta_t B_t \right\rangle_{l^2} \\
&= \left\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\right\|_{l^2}^2 + 2\eta_t\left\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}^t, B_t\right\rangle_{l^2} + \eta_t^2\left\|B_t\right\|_{l^2}^2
\end{aligned} \tag{15}$$

$$\begin{aligned}
&\left\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}^t, B_t \right\rangle \\
&= \left\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}^t, \frac{1}{\omega}\sum_{i=1}^{\omega}(\boldsymbol{\theta}^T\boldsymbol{K}[x_i] - [y_i])\boldsymbol{K}[x_i] \right\rangle \\
&+ \lambda\left\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}^t, \boldsymbol{\theta}^t \right\rangle \\
&= \left\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}^t, \frac{1}{\omega}\sum_{i=1}^{\omega}(\boldsymbol{\theta}^T\boldsymbol{K}[x_i] - [y_i])\boldsymbol{K}[x_i] \right\rangle \\
&+ \lambda\left\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}^t \right\rangle - \lambda\left\|\boldsymbol{\theta}^t\right\|_{l^2}^2 \\
&\leq \left\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}^t, \frac{1}{\omega}\sum_{i=1}^{\omega}(\boldsymbol{\theta}^T\boldsymbol{K}[x_i] - [y_i])\boldsymbol{K}[x_i] \right\rangle \\
&+ \frac{1}{2}\lambda\left\|\boldsymbol{\theta}^*\right\|_{l^2}^2 + \frac{1}{2}\lambda\left\|\boldsymbol{\theta}^t\right\|_{l^2}^2 - \lambda\left\|\boldsymbol{\theta}^t\right\|_{l^2}^2 \\
&\leq \left\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}^t, \frac{1}{\omega}\sum_{i=1}^{\omega}(\boldsymbol{\theta}^T\boldsymbol{K}[x_i] - [y_i])\boldsymbol{K}[x_i] \right\rangle \\
&+ \frac{\lambda}{2}\left\|\boldsymbol{\theta}^*\right\|_{l^2}^2 - \frac{\lambda}{2}\left\|\boldsymbol{\theta}^t\right\|_{l^2}^2
\end{aligned} \tag{16}$$

By the convexity, we get from

$$\begin{aligned}
&\left\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}^t, \frac{1}{\omega}\sum_{i=1}^{\omega}(\boldsymbol{\theta}^T\boldsymbol{K}[x_i] - [y_i])\boldsymbol{K}[x_i] \right\rangle \\
&\leq \frac{1}{2\omega}[\frac{1}{\omega}\sum_{i=1}^{\omega}(\boldsymbol{\theta}^*\boldsymbol{K}^T[x_i] - y_i)^2 \\
&- \sum_{i=1}^{\omega}(\boldsymbol{\theta}^t\boldsymbol{K}^T[x_i] - y_i)^2]
\end{aligned} \tag{17}$$

So (16) can turn into

$$\begin{aligned}
&\left\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}^t, B_t\right\rangle_{l^2} \leq \frac{1}{2}[\frac{1}{\omega}\sum_{i=1}^{\omega}(\boldsymbol{\theta}^*\boldsymbol{K}^T[x_i] - y_i)^2 \\
&+ \lambda\left\|\boldsymbol{\theta}^*\right\|_{l^2}^2] - \frac{1}{2}[\frac{1}{\omega}\sum_{i=1}^{\omega}(\boldsymbol{\theta}^t\boldsymbol{K}^T[x_i] - y_i)^2 + \lambda\left\|\boldsymbol{\theta}^t\right\|_{l^2}^2]
\end{aligned} \tag{18}$$

So we need to prove the correctness of this function

Lemma 3.3. $\lambda\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\right\|_{l^2}^2$
$\leq \{\frac{1}{\omega}\sum_{i=1}^{\omega}(\boldsymbol{\theta}\boldsymbol{K}^T[x_i] - y_i)^2 + \lambda\left\|\boldsymbol{\theta}\right\|_{l^2}^2\} -$
$\{\frac{1}{\omega}\sum_{i=1}^{\omega}(\boldsymbol{\theta}^*\boldsymbol{K}^T[x_i] - y_i)^2 + \lambda\left\|\boldsymbol{\theta}^*\right\|_{l^2}^2\}$

Proof.

$$\boldsymbol{\theta}^{\nu} = \nu\boldsymbol{\theta} + (1-\nu)\boldsymbol{\theta}^* \tag{19}$$

$$G(\nu) = \frac{1}{\omega}\sum_{i=1}^{\omega}(\boldsymbol{\theta}^{\nu T}\boldsymbol{K}[x_i] - y_i)^2 + \lambda\left\|\boldsymbol{\theta}^{\nu}\right\|_{l^2}^2 \tag{20}$$

$$\begin{aligned}
G'(\nu) =& 2\lambda(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T\boldsymbol{\theta}^{\nu} + \frac{2}{\omega}\sum_{i=1}^{\omega}(\boldsymbol{\theta}^T\boldsymbol{K}[x_i] - y_i)(\boldsymbol{\theta} \\
&- \boldsymbol{\theta}^*)^T\boldsymbol{K}[x_i]
\end{aligned} \tag{21}$$

Beacause $f'(\theta^*) = 0$,then

$$\begin{aligned}
&\lambda(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T\boldsymbol{\theta}^* + \frac{1}{\omega}\sum_{i=1}^{\omega}(\boldsymbol{\theta}^T\boldsymbol{K}[x_i] - y_i)(\boldsymbol{\theta} - \\
&\boldsymbol{\theta}^*)^T\boldsymbol{K}[x_i] = \boldsymbol{0}
\end{aligned} \tag{22}$$

$$
\begin{aligned}
G'(\nu) &= 2\lambda(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T\boldsymbol{\theta}^{\nu} - 2\lambda(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T\boldsymbol{\theta}^* \\
&\quad + \frac{2}{\omega}\sum_{i=1}^{\omega}(\boldsymbol{\theta}^{\nu T}\boldsymbol{K}[x_i]-y_i)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T\boldsymbol{K}[x_i] \\
&\quad - \frac{2}{\omega}\sum_{i=1}^{\omega}(\boldsymbol{\theta}^{*T}\boldsymbol{K}[x_i]-y_i)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T\boldsymbol{K}[x_i] \\
&= 2\lambda\nu(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T(\boldsymbol{\theta}-\boldsymbol{\theta}^*) + \frac{2\nu}{\omega}\sum_{i=1}^{\omega}((\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T\boldsymbol{K}[x_i])^2 \\
&\geq 2\lambda\nu(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T(\boldsymbol{\theta}-\boldsymbol{\theta}^*)
\end{aligned}
\tag{23}
$$

Observe that $G(1)-G(0)=\int_0^1 G'(\nu)\,d\nu$.Thus,

$$
\begin{aligned}
G(1)-G(0) &\geq 2\lambda(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\int_0^1 \nu\,d\nu \\
&= \lambda(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T(\boldsymbol{\theta}-\boldsymbol{\theta}^*)
\end{aligned}
\tag{24}
$$

The correctness of Lemma3.3 can be proved. □

Combining Lemma3.3 and (18),We can get

$$
\langle \boldsymbol{\theta}^*-\boldsymbol{\theta}^t, B_t\rangle \leq -\frac{\lambda}{2}\left\|\boldsymbol{\theta}^t-\boldsymbol{\theta}^*\right\|_{l^2}^2 \tag{25}
$$

Assume that k is the largest element of $\boldsymbol{K}[x_i]$, y is the largest element of $\boldsymbol{Y}$

Lemma 3.4. $\left\|\boldsymbol{\theta}^t\right\|_{l^2} \leq \dfrac{yk}{\lambda I}$

Proof. We utilize mathematical induction to prove it. When t = 1, the correctness of lemma 2.4 is obvious.
When $t=t_1$, we assume that $\left\|\boldsymbol{\theta}^{t_1}\right\|_{l^2} \leq \dfrac{yk}{\lambda l}$ is correct.

So When $t=t_1+1$,

$$
\begin{aligned}
\left\|\boldsymbol{\theta}^{t_1+1}\right\|_{l^2} &\leq (1-\lambda\eta_t)\left\|\boldsymbol{\theta}^t\right\|_{l^2} + \left\|\frac{\eta_t}{\omega}\textstyle\sum_{i=1}^{\omega} y_i K[x_i]\right\|_{l^2} \\
&\leq \frac{yk}{\lambda l} - \lambda\eta_t\frac{yk}{\lambda l} + \eta_t\frac{yk}{l} = \frac{yk}{\lambda l}
\end{aligned}
\tag{26}
$$

Above all, Lemma 2.4 can be proved.

□

Lemma 3.5.

$$
\left\|(\boldsymbol{\theta}^{tT}\boldsymbol{K}[x_i]-y_i)\boldsymbol{K}[x_i]\right\|_{l^2} \leq \frac{yk^3}{\lambda} + \sqrt{l}yk \tag{27}
$$

Proof.

$$
\begin{aligned}
\left\|(\boldsymbol{\theta}^{tT}\boldsymbol{K}[x_i]-y_i)\boldsymbol{K}[x_i]\right\|_{l^2} &\leq \left\|\boldsymbol{\theta}^{tT}\boldsymbol{K}[x_i]\right\|_{\infty}\left\|\boldsymbol{K}[x_i]\right\|_{l^2} \\
&\quad + y\left\|\boldsymbol{K}[x_i]\right\|_{l^2} \\
&\leq lk^2\left\|\boldsymbol{\theta}^t\right\|_{l^2} + y\sqrt{l}k \\
&\leq lk^2\frac{yk}{\lambda l} + y\sqrt{l}k \\
&= \frac{yk^3}{\lambda} + \sqrt{l}yk
\end{aligned}
\tag{28}
$$

□

Therefore,

$$
\begin{aligned}
\left\|B_t\right\|_{l^2}^2 &\leq \Big(\frac{1}{\omega}\sum_{i=1}^{\omega}\left\|(\boldsymbol{\theta}^T\boldsymbol{K}[x_i]-y_i)\boldsymbol{K}[x_i]\right\|_{l^2} + \lambda\left\|\boldsymbol{\theta}^t\right\|_{l^2}\Big)^2 \\
&\leq \Big(\frac{yk^3}{\lambda} + \sqrt{l}yk + \lambda\frac{yk}{\lambda l}\Big)^2
\end{aligned}
\tag{29}
$$

So,combing (15) (25) (29),we can get

$$
\begin{aligned}
\left\|\boldsymbol{\theta}^{t+1}-\boldsymbol{\theta}^*\right\|_{l^2}^2 &\leq (1-\lambda\eta_t)\left\|\boldsymbol{\theta}^t-\boldsymbol{\theta}^*\right\|_{l^2}^2 \\
&\quad + \eta_t^2\Big(\frac{yk^3}{\lambda} + \sqrt{l}yk + \lambda\frac{yk}{\lambda l}\Big)^2
\end{aligned}
\tag{30}
$$

So this algorithm is a linear convergence of Q-th.